Integration of Fuzzy Clustering Technique with Big Data for Disease Diagnosis

[1]Dr.S.Sapna, [2]M. Pravin Kumar
[1]Assistant Professor, Department of Computer Science, Bharathidasan College of Arts & Science, Erode, INDIA
[2]Assistant Professor, Department of ECE, KSR College of Engineering, Tiruchengode, INDIA

***Abstract*—This paper presents an integrative approach to predict the diabetic disease from clinical big data. The clinical database is generally redundant, incomplete, vague and unpredictable. A fuzzy technique is integrated to handle vagueness in clinical big data. A compact fuzzy model is created using subtractive clustering. The main aim of proposed approach is to increase the prediction accuracy. The outcomes indicate that the proposed method can be used effectively in healthcare to diagnose the diabetic disease.**

Keywords—big data, clinical, fuzzy, hidden knowledge, prediction, subtractive clustering.

I. INTRODUCTION

The main aim of clinical informatics is to take in real world medical data from all stages of human being to help improve our understanding of treatment and medical practice. The volume of clinical data is likely to increase drastically in the forthcoming years. Health informatics tools include amongst others computers, clinical guidelines, formal medical terminologies, and information and communication systems [10,9]. Healthcare reimbursement models are changing; meaningful use and pay for performance are emerging as critical new factors in today's healthcare environment. Although profit is not and should not be a primary motivator, it is vitally important for healthcare organizations to acquire the available tools, infrastructure, and techniques to leverage big data effectively [8,16]. The field of Health Informatics is on the crossover of its most exciting period to date, entering a new era where technology is starting to handle Big Data, bringing about unlimited potential for information growth [11,6]. The large volume of data in health care exceeded our human ability for diagnosing disease without powerful tools. Big Data analytics and Fuzzy logic techniques, helps to better understand the objectives of diagnosing and treating patients in need of healthcare.

Big data in healthcare refers to electronic health data sets so large and complex that they are difficult (or impossible) to manage with traditional software and/or hardware; nor can they be easily managed with traditional or common data management tools and methods. Big data in healthcare is overwhelming not only because of its volume but also because of the diversity of data types and the speed at which it must be managed as it includes clinical data and clinical decision support systems [12]. Fuzzy logic system is used for solving a wide range of problems in different application involving large volume of data. It also allows us to introduce the learning and adaptation capabilities. The fuzzy set framework has been used in several different process of diagnosis of disease. Fuzzy logic is a computational paradigm that provides a mathematical tool for dealing with the uncertainty and the imprecision typical of human reasoning [14].

The clinical data are usually uncertain and vague to arrive at a conclusion. So, in this paper fuzzy subtractive clustering technique is proposed to diagnose the diabetic disease from the clinical diabetic data. This integrative approach provides flexible information capability for medical practitioners in handling ambiguous situations.

II. BIG DATA

Big Data is data whose scale, diversity, and complexity require novel architecture, methods, procedures, and analytics to manage it and extract value and hidden knowledge from it. The greatest challenges is to deal with large dataset with high amount of dimensionality, together in terms of the number of features the data has, as well the number of rows of data that user is dealing with. Big Data requires different approaches with an aim to solve new problems or existing problems in a better way.

International Business Machines [3] estimates that every day user creates 2.5 quintillion bytes of data, so much that 90% of the data in the world today has been created in the last two years alone. The big data come from sensors that are used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals.

Data is a raw unorganized facts, is in and of itself worthless. Information means potential valuable concepts based on data. Knowledge is what we understand based upon information. Wisdom means effective use of knowledge in decision making. The key enablers for the appearance and growth of Big-Data are increase in storage capabilities, increase in processing power and availability of data. There are huge volumes of data in the world. From the beginning of recorded time until 2003, User's created 5 billion gigabytes

(109) of data. In 2011, the same amount was created every two days. In 2013, the same amount of data is created every 10 minutes. Prior to 2012 the US was the largest single contributor to global data. 32 % united states, 13% china, 4% India 19% Western Europe and 32% of rest of world. In 2020 the emerging markets are showing the largest increases in data growth 23% United States, 21% china, 6% India, 15% Western Europe and 35% rest of world. In 2012 the amount of information stored worldwide exceeded 2.8 zetabytes (1021).

By 2016 the cumulative size of all the world's data center's excepted to exceed to 16,000 acres which is equivalent in area to a two lane highway stretching from Tokyo to San Francisco over 5,000 miles. An estimated 33% of information could be useful if appropriately tagged and analysed. The amount actually analysed is only 0.5%. By 2020 the total amount of data stored is expected to be 50x larger than today. In the digital universe all that has been created and stored are nearly half of which is unprotected.

A. V's of Big Data

The five V's of big data are volume, velocity, variety and veracity. Volume means data at rest. Enterprises are packed with ever-growing data of all types, easily amassing terabytes (1012) - even petabytes (1015) - of information. Turn 12 terabytes of Tweets created each day into improved product sentiment analysis. Convert 350 billion annual meter readings to better predict power consumption. Velocity means data in motion. Sometimes 2 minutes is too late. For time-sensitive processes such as catching fraud, big data must be used as it streams into the enterprise in order to maximize its value. Scrutinize 5 million trade events created each day to identify potential fraud. Analyze 500 million daily call detail records in real-time to predict customer churn faster. The latest is 10 nano seconds delay is too much. Variety means data in many forms i.e., structured and unstructured data such as text, sensor data, audio, video, click streams, log files and more. New insights are found when analyzing these data types together. Monitor 100's of live video feeds from surveillance cameras to target points of interest. Exploit the 80% data growth in images, video and documents to improve customer satisfaction. Veracity means data in doubt that is uncertainty due to data inconsistency and incompleteness, ambiguities, latency, deception, Model approximations. Value means a clear understanding of costs and benefits. The challenge is to find a way to transform raw data into information that has value, either internally, or for making a business out of it. The pictorial representation of V's of Big data is shown below in the Fig 1.

Big Data is generated by social media and networks (all of us are generating data), Scientific instruments (collecting all sorts of data), Mobile devices (tracking all objects all the time), Sensor technology and networks (measuring all kinds of data). The progress and innovation is no longer hindered by the ability to collect data. But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion would certainly be the challenges of information technology community [3,13].

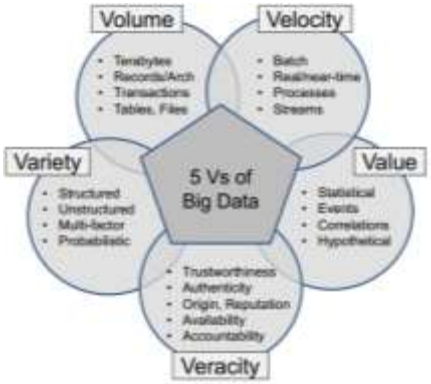

Fig. 1 V's of Big Data

III. DIABETES MELLITUS

Diabetes Mellitus often simply referred to as diabetes is the coercive effect of insulin on the glucose metabolism. Insulin is a hormone central to regulating carbohydrate and fat metabolism in the body. Insulin is produced from the islets of langerhans [1,14]. In Latin the word Insula means-"island". Its concentration has wide spread effect throughout the body. When control of insulin levels fails, diabetes mellitus will result. As a consequence, insulin is used medically to treat some forms of diabetes mellitus. Diabetes type 1 is lack of it whereas diabetes type 2 is the resistance towards it. Not only insulin regulates the glucose in the blood but it is also responsible for lipid metabolism. Insulin Secretion from beta-cells is principally regulated by plasma glucose levels [14,15]. Increased uptake of glucose by pancreatic beta-cells leads to a concomitant increase in metabolism. One must understand that insulin is offered as medicine only when the above criteria are broken. Physicians will become familiar with other aspects of managing the patient with diabetes, including the importance of postprandial glucose control, diabetes self-management training etc.

Most of the food that is eaten is converted to glucose, or sugar which is used for energy. The pancreas secretes insulin which carries glucose into the cells of our bodies, which in turn produces energy for the perfect functioning of the body. When you have diabetes, your body either doesn't make enough insulin or cannot use its own insulin as well as it should [5,14]. This causes sugar to build up in your blood leading to complications like heart disease, stroke, neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage and death.

A. General Symptoms of Diabetes

Increased thirst, Increased urination - Weight loss, Increased appetite – Fatigue, Nausea and/or vomiting - Blurred vision, Slow-healing infections - Impotence in men.

B. Types of Diabetes

Type 1: Diabetes also called as Insulin Dependent Diabetes Mellitus **(IDDM)**, or Juvenile Onset Diabetes Mellitus commonly seen in children and young adults however, older patients do present with this form of diabetes on occasion. In type 1 diabetes, the pancreas undergoes an autoimmune attack by the body itself therefore; pancreas does not produce the hormone insulin. The body does not properly metabolize food resulting in high blood sugar (glucose) and the patient must rely on insulin shots. Type I disorder appears in people younger than 35, usually from the ages 10 to 16.

Type II: Diabetes is also called as Non-Insulin Dependent Diabetes Mellitus (NIDDM) or Adult Onset Diabetes Mellitus. Patients produce adequate insulin but the body cannot make use of it as there is a lack of sensitivity to insulin by the cells of the body. Type II disorder occurs mostly after the age 40.

Gestational Diabetes: Diabetes can occur temporarily during Pregnancy called as Gestational Diabetes which is due to the hormonal changes and usually begins in the fifth or sixth month of pregnancy (between the 24th and 28th weeks). It usually resolves once the baby is born. 25-50% of women with eventually develop diabetes later in life, especially in those who require insulin during pregnancy and those who are overweight after their delivery.

C. Diagnostic Tests

Urine Test: A urine analysis may be used to look for glucose and ketones from the breakdown of fat. However, a urine test alone does not diagnose diabetes. The following blood glucose tests are used to diagnose diabetes.

Fasting Plasma Glucose Level (FPG): The normal range of fasting blood glucose is <100 mg/dl. It is done after 8-12 hours of fasting. People with fasting glucose levels from 100-125 mg/dl are considered to have impaired fasting glucose. Patients with FPG >126 are consider to have diabetes mellitus.

Post Prandial Plasma Glucose Level (PPG): A blood sugar test taken after two hours of a meal is known as the post prandial glucose test or PPG. The normal range for PPG is <140 mg/dl. People with fasting glucose levels from 140-200 mg/dl are considered to have impaired glucose tolerance. Patients with PPG >200 mg/dl are consider to have diabetes mellitus [14].

IV. CLUSTERING TECHNIQUE

Clustering of numerical data forms the basis of many classification and system modeling algorithms. The purpose of clustering is to identify natural groupings of data from a large data set to produce a concise representation of a system's behavior. The idea of data grouping, or clustering, is simple in its nature and is close to the human way of thinking; whenever there are large amount of data are presented, it usually tend to summarize the huge number of data into a small number of groups or categories in order to further facilitate its analysis. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis, information retrieval, and bioinformatics [7].

A. Need of Clustering

In the fuzzy logic system if the number of inputs to the system is increased, then the number of rules increases exponentially. In predicting diabetes, five inputs and two outputs are considered. So it is difficult to proceed in the system and the concept of clustering is followed. In the clustering method the cluster centers are determined where each cluster center belongs to one rule in fuzzy logic.

B. Subtractive Clustering

The subtractive clustering method assumes each data point is a potential cluster center and calculates a measure of the likelihood that each data point would define the cluster center, based on the density of surrounding data points. For subtractive clustering method proposed by [2,4], data points have to be rescaled to [0, 1] in each dimension. Each data point $z_j = (x_j, y_j)$ is assigned potential P_j, according to its location to all other data points.

$$P_i^* = \sum_{j=1}^{n} e^{-\alpha \left| x^i - x^j \right|^2} \quad \ldots (1)$$

Where

$$\alpha = \frac{\gamma}{r_a} \quad \ldots (2)$$

P_i^* is the potential value i data as a cluster center
α is the weight between i th data to j th data
x is the data point
γ is the variables(commonly set 4)
r_a is a positive constant called cluster radius

The potential of a data point to be a cluster center is higher when more data points are closer. The data point with the highest potential, denoted by P_i^* is considered as the first cluster center $c_1 = (d_1, e_1)$. The potential is then recalculated for all other points excluding the influence of the first cluster center according to the Equation (3).

$$P_i^* = P_i^* - P_K^* \zeta \quad \ldots (3)$$

Where

$$\zeta = e^{-\beta \left\| x^r - c^k \right\|^2} \quad \ldots (4)$$

$$\beta = \frac{4}{r_b^2} \quad \ldots (5)$$

$$r_b = r_a * \eta \quad \ldots (6)$$

P_i^* is the new potential value i-data

P_k^* is the potential value data as cluster center

r_b is a positive constant

c is the cluster center of data

β is the weight of I data to cluster center

r_i is the distance between cluster center

η is the quash factor

Again, the data point with the highest potential P_k^* is considered to be the next cluster center c_k, if

$$\frac{d_{min}}{r_a} + \frac{P_k^*}{P_I^*} \geq 1 \qquad \ldots (7)$$

with d_{min} is the minimal distance between C_1 and all previously found cluster centers, the data point is still accepted as the next cluster center c_1. Further iterations can then be performed to obtain new cluster centers c_2. If a possible cluster center does not fulfill the above described conditions, it is rejected as a cluster center and its potential is set to 0. The data point with the next highest potential P_K^* is selected as the new possible cluster center and re-tested. The clustering ends if the following condition is fulfilled by Equation (8).

$$P_K^* < \varepsilon P_i^* \qquad \ldots (8)$$

Where ε is the reject ratio.

Indicative parameters values for r_a, η, ε and ε^* have been suggested. Each cluster center is considered as a fuzzy rule that describes the system behavior of the distance to the defined cluster centers.

$$\mu_j^{ik} = e^{-\alpha \left\| x_j^k - c_j^k \right\|^2} \qquad \ldots (9)$$

Equation (9) is a common form of subtractive clustering, hence it needs to create an algorithm to process data clustering. This paper proposes an algorithm to cluster the clinical diabetic data to predict the disease effectively.

V. IMPLEMENTATION OF SUBTRACTIVE CLUSTERING FOR PREDICTION OF DIABETES MELLITUS

The clinical diabetic data were collected from various diabetic centers at Erode, Tamilnadu. About 1050 diabetic patients' data were considered for this prediction and some of which is shown in Table 1. The subtractive clustering method is applied to the input-output of the clinical diabetic data to analyze their performance. The inputs considered are Age, FPG-Fasting Plasma Glucose, PPG-Post Prandial Plasma Glucose, G-Gender, P/NP-Pregnant or Non Pregnant and the outputs are D-Diabetic Status and T1/T2/GD (T1 – Type 1/ T2 – Type 2/GD – Gestational Diabetes) [14]. As five inputs and two outputs are considered, the number of rules will increase exponentially, so the subtractive clustering technique is applied to get the optimal number of rules. By applying subtractive clustering the number of cluster centers is obtained, where each cluster center belongs to one rule in fuzzy logic and also the hidden knowledge from the clinical database is predicted which aids the physician in decision-making. This prediction system also helps the user to anticipate by himself / herself whether he/she is affected with diabetes or not and also to which type of diabetes he/she belongs.

TABLE. 1 PRACTICALLY OBSERVED CLINICAL DATABASE OF DIABETES

INPUTS						OUTPUT	
S.No.	**Age**	**FPG (mg/dl)**	**PPG (mg/dl)**	**G**	**P / NP**	**D**	**T1/ T2 / GD**
1	52	95	290	0	0	0.7	0.7
2	51	109	452	1	0	0.91	0.7
.	.	.	.	.	.	.	.
1050	43	178	99	0	1	0.68	0.8

A. Results and Discussions

The subtractive clustering technique is applied to clinical diabetic database given in Table 1, and its performance obtained using Matlab R2007b is shown in Fig 2. After clustering the data, subtractive clustering produces 9 cluster centers, where each cluster center belongs to one rule in fuzzy logic and the root mean square error is found to be 0.9846. This means if each cluster center is equal to one rule, there are only nine rules to represent 1050 data as shown in Fig 3.

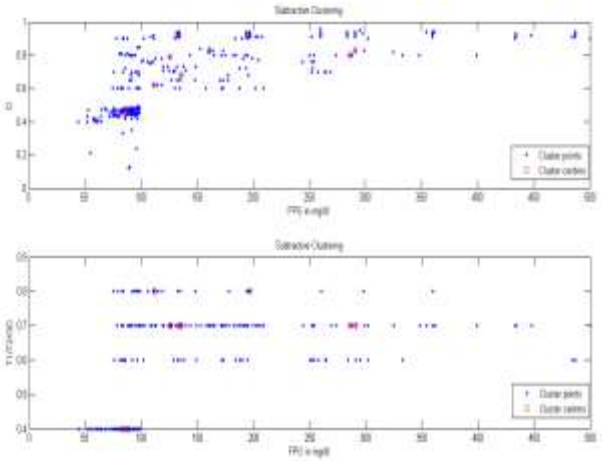

Fig. 2 Performance of Subtractive Clustering for Clinical Diabetic Database

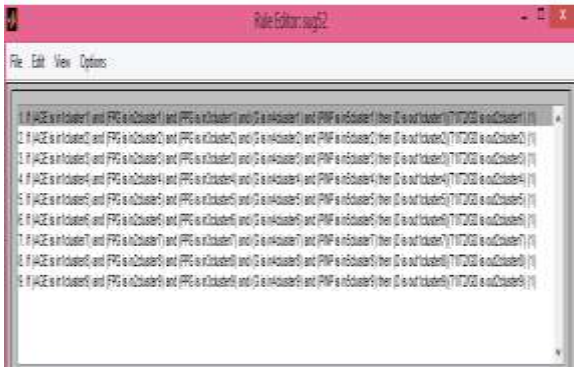

Fig. 3 Fuzzy If-Then Rules

The Fuzzy Inference System (FIS) obtained using subtractive clustering for clinical diabetic database is shown in Fig 4. In FIS the input and output attributes are specified. Five input and two output system are built using FIS for clinical diabetic database.

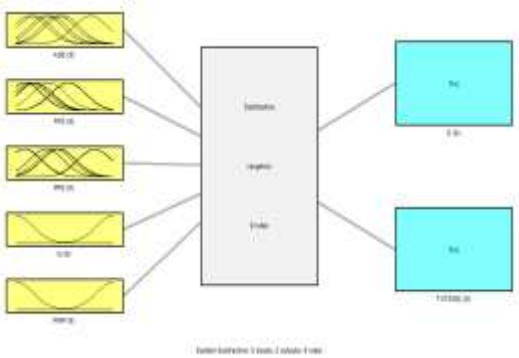

Fig. 4 Fuzzy Inference System

Fig. 5 Rule View

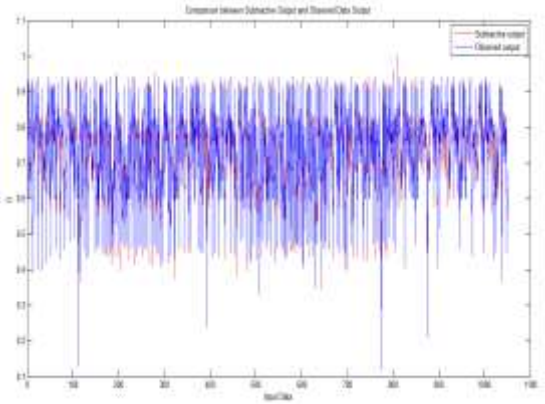

Fig. 6 Comparison between Subtractive Output and Observed Data Output

The rule view obtained using subtractive clustering for clinical diabetic database is shown in Fig 5. The rule viewer displays a roadmap of the whole fuzzy inference process. The system readjusts and computes the new output when the red line indices are moved around the corresponding input attributes. The clustering process stops when the maximum number of iterations is reached. For the implemented practical diabetic data the clustering process stops at 80^{th} iteration. The comparison between subtractive clustering output and observed data output for D (Diabetes) is shown in Fig 6. From Fig 6, it is observed that subtractive clustering is close to observed output. From Fig 5, the hidden knowledge from the clinical database can be predicted which aids the physician in decision-making. This prediction system also helps the user to anticipate by himself / herself whether he/she is affected with diabetes or not and also predicts to which type of diabetes the he/she belongs.

VI. CONCLUSION

Clinical processes progress over time and all the clinical actions are indescribable without considering time. Therefore time is vital to many medical domain problems. For the collected clinical diabetic data the subtractive clustering method is applied and its performance is analyzed. Subtractive clustering technique reduces the number of rules in the rule base by eliminating the redundant rules thereby reducing the computational time. The proposed integrated approach helps the physician and medical experts to diagnosis the diabetic disease effectively.

REFERENCES

[1] Aarogyam Preventive Health Care Magazine: Thyrocare's, Basics of diabetes, Vol. 10, no. 6, 2006.
[2] Agus Priyono, Ahmad Jais Alias and Muhammad Ridwan, "Generation of fuzzy rules with subtractive clustering", Journal Technology, Vol. 43, no. D, 2005, pp. 143-153.

[3] Bringing big data to the enterprise #ibmbigdata website, (2012), Available from: http://www-01.ibm.com/software/data/bigdata/what-is-big-data.html.

[4] Chiu .S.L., "Fuzzy model identification based on cluster estimation", Journal of Intelligent & Fuzzy Systems, Vol. 2, 1994, pp. 267-278.

[5] Diabetes and YOU your guide to living well with diabetes, Novo Nordisk, LEAD GROUP.

[6] HongboZoua, YongenYub, WeiTangc, Hsuan-Wei MichelleChend, "FlexAnalytics: A flexible data analytics framework for Big Data Applications with I/O performance improvement", Big Data Research, Elsevier Publications, Vol. 1, 2014, pp. 4-13.

[7] Jang J.S.R., Mizutani E. and Sun C.T., "Neuro Fuzzy and Soft Computing", Prentice-Hall, Englewood Cliff, New Jersey, 1997.

[8] LaValle S, Lesser E, Shockley R, Hopkins MS, Kruschwitz N: "Big data, analytics and the path from insights to value", MIT Sloan Manag Rev 2011, Vol. 52, no.2, pp. 20–32.

[9] Mettler T, Raptis DA (2012). "What constitutes the field of health information systems? Fostering a systematic framework and research agenda". Health Informatics Journal Vol. 18, no. 2, pp. 147– 56 .

[10] Matthew Herland, Taghi M Khoshgoftaar and Randall Wald, "A review of data mining using big data in health informatics", journal of Big Data, a Springer open journal, Vol. 1, no. 2, 2014 pp. 1-35.

[11] Peter Augustine D, "Leveraging Big Data analytics and Hadoop in developing India's healthcare services", International Journal of Computer Applications, Vol. 89, no. 16, 2014, pp. 44-50.

[12] Priyanka K and Nagarathna Kulennavar, "A survey on big data analytics in health care", IJCSIT, International Journal of Computer Science and Information Technologies, Vol. 5, no. 4, 2014, pp.5865-5868.

[13] Pooja, Sandeep Jaglan, Reema Gupta, "Big data: advancement in data analytics", International Journal of Computer Technology & Applications, Vol. 5, no. 4, 2014, pp.1466-1469.

[14] Sapna S and Pravin Kumar M., "Prediction of uncertainty in clinical database using clustering technique", International Journal of Innovative Research in Technology, Vol. 1, no. 10, 2015, pp. 98-102.

[15] The New Indian Express, Health Article Tue, Aug 14, 2007, by Dr. K.Bhujang Shetty, pp. 1.

[16] Wullianallur Raghupathi and Viju Raghupathi, "Big Data analytics in healthcare: promise and potential", Raghupathi and Raghupathi Health Information Science and Systems, Vol. 2, no. 3, 2014, pp 2-10.